A bioinformatic web server to cut protein structures in terms of Protein Units.

Jean-Christophe Gelly[#] & Alexandre G. de Brevern[#]*

INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot – Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.

Mails : jean-christophe.gelly@univ-paris-diderot.fr, alexandre.debrevern@univ-paris-diderot.fr

* Corresponding author:

Mailing address: Dr. de Alexandre G. de Brevern, INSERM UMR-S 665, DSIMB, Université Paris Diderot – Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France

[#] The authors wish it to be known that, in their opinion, the first and the last authors should be regarded as joint First Authors.

Abstract

Analysis of the architecture and organization of protein structures is a major challenge to better understand protein flexibility, folding, functions and interactions with their partners and to design new drugs.

Protein structures are often described as series of α-helices and β-sheets, or at a higher level as an arrangement of protein domains. Due to the lack of an intermediate vision which could give a good understanding and description of protein structure architecture, we have proposed a novel intermediate view, the Protein Units (PUs). They are novel level of protein structure description between secondary structures and domains. A PU is defined as a compact sub-region of the 3D structure corresponding to one sequence fragment, defined by a high number of intra-PU contacts and a low number of inter-PU contacts. The methodology to obtain PUs from the protein structures is named Protein Peeling (PP). For the algorithm, the protein structures are described as a succession of Cα. The distances between Cα are translated into contact probabilities using a logistic function. Protein Peeling only uses this contact probability matrix. An optimization procedure, based on the Matthews' coefficient correlation (MCC) between contacts probability sub matrices, defines optimal cutting points that separate the region examined into two or three PUs. The process is iterated until the compactness of the resulting PUs reaches a given limit. An index assesses the compactness quality and relative independence of each PU.

Protein Peeling is a tool to better understand and analyze the organization of protein structures. We have developed a dedicated bioinformatic web server: Protein Peeling 2 (PP2). Given the 3D coordinates of a protein, it proposes an automatic identification of protein units (PUs). The interface component consists of a web page (HTML) and common gateway interface (CGI). The user can set many parameters and upload a given structure in PDB file format to a perl core instance. This last component is a module that embeds all the

information necessary for two others softwares (mainly coded in C to perform most of the computation tasks and R for the analysis). Results are given both textually and graphically using JMol applet and PyMol software. The server can be accessed from http://www.dsimb.inserm.fr/dsimb_tools/peeling/. Only one equivalent on line methodology is available.

Introduction

The proteins are a succession of amino acids joined together by peptide bonds. They are crucial macromolecules implicated in major physiological processes and also most of the diseases. Since the elucidation of the first protein structure by Max Perutz and John Kendrew in the late 50's (Kendrew et al. 1958; Perutz et al. 1960), it was experimentally demonstrated that functional proteins adopt a three dimensional structure (3D) defined by the spatial arrangement of the atoms of its amino acids. Protein 3D structures are still resolved using X-ray crystallography since the last 50 years (Kendrew et al. 1958). The process by which a protein adopts this three-dimensional structure under natural condition from an initial disordered state is called folding. Native protein structures are maintained by inter-residue interactions. Anfinsen demonstrated that the amino acid sequence alone contain all the information needed to obtain a functional protein structure (Anfinsen et al. 1961). Otherwise molecular mechanism responsible for this self-assembly is poorly understood and remains one of the most fundamental problems in biological sciences.

From the beginning of biochemical sciences, interesting characteristics have been determined or theoretically predicted. Thus, some amino acids favor to adopt local structures called repetitive secondary structures: α-helix and β-sheet due to their physicochemical properties. The combination of theses secondary structures elements and other non-regular form the final structure of the protein and characterize a particular and functional protein

topology (Richardson 1981).

Various studies had also shown that protein structure fold can be represented into units called protein structural domains. Proteins can be constituted by one unique domain while others are combinations of many ones. Domains represent not only structural meaningful elements but also facilitate the understanding of protein architecture. Quasi structural independence is the major characteristic of domains. Sometimes theses domains had also a well defined function and were evolutionary conserved (Ponting and Russell 2002). One aim is to simplify analysis into more significant component based on geometric and physicochemical properties. Indeed great part of protein domains are organized around a hydrophobic core and some are able to fold independently and exhibit a well defined topology. More than one thousand different domains have been identified in structural databases, *e.g.*, SCOP (Murzin et al. 1995), CATH (Orengo et al. 1997) or FSSP (Holm and Sander 1997). Defining automatic procedures for reliable domain assignment is an essential task for the generation of pertinent domain databases used for relevant scientific studies (Heger et al. 2005). The main idea behind these approaches is that the inter-domain interaction is weaker than the intra-domain interaction.

Despite availability of these various tools, it remains hard to describe and understand protein structures diversity with theses methodologies. A clear gap exists between an elementary view of protein structure as a succession of secondary structure elements, and a more complex view of protein structural domains. It lacks a level of description complementary between secondary structures and domains, a kind of intermediate view of structural organization and complexity. Some authors have proposed such a supplementary level, consistent with folding models. Wetlaufer was the first to examine the organization of known structures and suggested that the early stages of 3D structure formation, *i.e.*, nucleation, occur independently in separate parts of these molecules (Wetlaufer 1973; 1981).

These folding units have been proposed to fold independently during the folding process, creating structural modules which give birth to the native structure.

Protein domains identification methods. Protein structures can be seen as composed of single or multiple functional domains that can fold and function independently. Dividing a protein into domains is useful for more accurate structure and function determination and explanation of folding process. Automatic domain parsing is based on a common simple principle: inter-domain interaction is weaker than the intra-domain interaction and intra-domain are strong enough to maintain stability (Wetlaufer 1973; Rossman and Liljas 1974; Richardson 1981; Wetlaufer 1981). Domain definition is simply the result of this cutting process. Many different procedures to assign protein structural domains have been developed. DETECTIVE method is based on the idea that domains have a hydrophobic interior (Swindells 1995), while Wodak and Janin used an iterative approach based on surface areas with an iterative cleavage of the native structure (Wodak and Janin 1981). Gaussian network model could also be used (Kundu et al. 2004). Different algorithms to hierarchically split proteins into compact units have been proposed (Lesk and Rose 1981; Wodak and Janin 1981; Sowdhamini and Blundell 1995; Swindells 1995; Tsai and Nussinov 1997; Kundu et al. 2004). Their goal was to describe protein structure at different organization levels (Taylor 2007). Nonetheless, the problem of dividing a protein structure into domains is not yet solved.

PUU (Holm and Sander 1994) is a recursive top-down approach which uses a hypothetical model of autonomously folding units corresponding to protein domains. A hierarchical 5-level filtering process is applied during partitioning of the structure, it tries to conserve long length protein fragment, cut flexible regions and not secondary structure. DomainParser uses a top-down approach to domain decomposition implemented using a graph theoretical approach (Xu et al. 2000; Guo et al. 2003). Its main problem is the failure to continue successful

partitioning. Protein Domain Parser (Alexandrov and Shindyalov 2003) is based on the assumption that the expected number of contacts between two domains depends on their surface areas. DOMAK (Siddiqui and Barton 1995), 3Dee (Dengler et al. 2001; Siddiqui et al. 2001), DETECTIVE (Swindells 1995), DALI (Holm and Sander 1998), STRUDL (Wernisch et al. 1999), and DDOMAIN (Zhou et al. 2007)) are build on similar approaches. Interestingly, they are often benchmarked on a manual definition of structural domains (Joshi 2007) as SCOP (Murzin et al. 1995). The difficulty of defining automatically structural domains has been often been shown, *e.g.*, (Holland et al. 2006) and no consensus could be easily found. An important point is the size of protein domains which always remains important (often more than a hundred residues) and so does not reflect protein folding early steps.

Small compact unit identification methods. Thus, many researchers have tried to determined smaller protein units which could represent earlier event of the protein folding and the smallest basic element of structure organization. The most common view was to define a hierarchically splitting of proteins into compact units (Go 1981; Lesk and Rose 1981; Wodak and Janin 1981; Janin and Wodak 1983; Sowdhamini and Blundell 1995; Tsai and Nussinov 1997; Guo et al. 2003; Pugalenthi et al. 2005). During the 70's Wetlaufer proposed that the early stages of 3D structure formation, *i.e.*, nucleation, occur independently in separate parts of these molecules (Wetlaufer 1973; 1981). These folding units have been proposed to fold independently during the folding process, creating structural modules which can be assembled to give the native structure.

Go identifies basic structural unit by Cα-Cα distance map and visual inspection of protein structures (Go 1981) while Janin and Wodak algorithm search along polypeptide the partitioning point generating units with the smallest surface interaction (Janin and Wodak

1983). Folding unit as defined by Lesk & Rose (Lesk and Rose 1981) are selected by a bottom-up hierarchical approach using inertial ellipsoids minimal area of small fragments. Later many methods based on different principles have been proposed. One of the most recent methodology is DIAL (Sowdhamini and Blundell 1995; Pugalenthi et al. 2005) and his database (Sowdhamini et al. 1996). DIAL algorithm determined small compact unit by hierarchical approach based on distances between secondary structure elements.

Protein Unit and Protein Peeling. We have likewise developed a method called Protein Peeling (Gelly et al. 2006a). This algorithm dissects a protein into Protein Units (PUs). A PU is a compact sub-region of the 3D structure corresponding to one sequence fragment. The basic principle is that each PU must have a high number of intra-PU contacts, and, a low number of inter-PU contacts. Thus, organization of protein structures can be considered in a hierarchical manner: secondary structures are the smallest elements, and, Protein Units are intermediate elements leading to structural domains.

Bioinformatics methodology needs to be widely available and distributed to be useful for the scientific community. A web server (http://www.dsimb.inserm.fr/dsimb_tools/peeling/) dedicated to Protein Peeling, has been developed for this purpose. It is now the second improved version of our approach (Gelly et al. 2006b).

Methodology of Protein Peeling

Protein Peeling algorithm works from the Cα-contact matrix translated into contact probabilities. A PU (or the protein at the beginning) is associated to a protein sequence ***s***

comprised between positions [*i, j*], ($i < j$, i=1 and j=N at the beginning).The sequence is cut into two parts $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ associated with the positions [*i*, *m*] and [*m* + 1, *j*] respectively. The symmetric contact probability sub-matrix associated to the sequence $\boldsymbol{s}$ is shared into 3 sub-matrices corresponding to the sum of the contact probabilities between the residues of $\boldsymbol{s_1}$ with itself (noted A), $\boldsymbol{s_2}$ with itself (B), and, $\boldsymbol{s_1}$ with $\boldsymbol{s_2}$ (C). To assess the presence of numerous contacts within the sub-units $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ and a limited number of contacts between them, Matthews' coefficient correlation (*MCC*) is used (Matthews 1975). The *MCC* measure is translated into a *partition index*, $PI_{ij}(m)$:

$$PI_{i,j}(m) = \frac{AB - C^2}{(A + C)(B + C)}$$

Thus, the quality of the splitting of the PU into two sub-units is quantified via a correlation. The complete absence of contacts between these two sub-units (*i.e.*, $C = 0$) leads to a maximal value of the partition index (*i.e.*, 1). A large presence of contacts between sub-units ($C > 0$) induces a low *PI* value. The cutting process cuts in 2 or 3PUs. To characterize the compactness of PUs defined, a compaction index (*CI*) based on mutual information is calculated (Etchebest et al. 2005; Hazout 2007)., it uses the sum of the probabilities associated with each PUs and indicates when to stop cutting, when it reaches a given threshold *R* (see (Gelly et al. 2006a) for more details and especially the Figure 1).

The process is recursively done. It is iterated until the compactness of the resulting PUs reaches a given limit, fixed by the user. A refinement of cutting is carried out thanks to the method of *pruning* which checks that PUs lately generated are compact (Gelly et al. 2006b).

Comparable approaches

The only comparable method is so DIAL. The approach is available at http://caps.ncbs.res.in/DIAL/. It considers small units as clusters of secondary structure elements. In a first step, α-helices and β-strands are first clustered using inter-secondary structural distances between Cα positions. In a second step, dendograms based on this distance measure are used to identify sub-domains. Their goal was to describe the different levels of protein structure organization.

Figure 1 shows the use of DIAL Web server to cut the structure of dialkylglycine decarboxylase (PDB code 1ZOB (Berman et al. 2000)). Figure 1a shows the website, Figure 1b is the result of cutting by DIAL of dialkylglycine decarboxylase. For this particular protein, two regions are found. They are represented on the structure on Figure 1c. It is a nice static view of the protein using molscript software (Kraulis 1991).

Protein Peeling web server

The flowchart representation of Protein Peeling web server is shown on Figure 2. Different languages and softwares are used. The web page in HTML shows on the upper left is the entrance point of PP2 webserver (see Figure 3, http://www.dsimb.inserm.fr/dsimb_tools/peeling/). After the submission of PDB file by the user, the common gateway interface (CGI) gets the values from the web form and transmitting it to the perl core instance. Then PDB file undergoes appropriate treatment. It start by the cleaning of the PDB files to ensure a correct format and afterward by the launching of secondary structure assignment done by DSSP software (Kabsch and Sander 1983). Then, the perl module launches the protein peeling main software (done in C language

for computational efficiency), that reads the clean PDB file, the secondary structure assignment and the different options. The protein peeling process is done and compactness indices are computed.

In a second step, render programs perform visualization of results. R software scripts (Ihaka and Gentleman 1996) are dedicated to visualize (i) the hierarchical peeling of the protein structures, (ii) the probabilities contact matrix and (iii) schematic representation of PUs in sequence with their contents in secondary structures. Two visualization softwares are used:

(i) Ray tracing proteins structures relies on PyMol (DeLano 2002) which gives excellent rendering. The perl core creates a dedicated PyMol script which is used and is also given to the user which can adapt to its own needs. The format conversion and the post-rendering of the pictures was managed by ImageMagick suite (ImageMagick).

(ii) An interactive visualization is also possible through the JMol applet (JMol) which is based on a Java Virtual Machine.

The perl core generates finally a complete web page (see Figure 2, left) that summarizes all the output information.

Example of protein cutting

Figures 4 and 5 show the cutting of dialkylglycine decarboxylase through Protein Peeling approach (PDB code 1ZOB). Figure 4b shows the dendogram obtained with default parameters; the cutting is so quite impressive. For R^2 equals to 20, a first event appears, it cuts the protein into two much misbalanced PUs (1-26 and 27-431). In a recent study (Faure et al. 2009), we have shown that Protein Peeling can detect mobile extremities. These last have

fewer constraints than the hydrophobic core of the protein and so are often considered as "mobile" (Jacob and Unger 2007). Our "mobile" extremities have been detected as PU, representing less than 20% of the size of protein which are cut early in the process of peeling and is not cut again. Half of the proteins have been detected associated to mobile extremities. Here, our mobile N-terminus (residue 1-26) is mainly helical which the case is for 2/3 of N-termini. As α-helices are not conditioned by long range contacts within the sequence like β-sheets; this tendency seems logical. Its *CI* value is low (0.26).

The second cutting event is for R^2 equals to 70, the splitting event is not a simple dual one, but three PUs are generated, with one short (27-59) and two longer PUs (60-326 and 327-431). The first one mainly composed of β-strands will not be cut again as the last one which is a bundle of 3 α-helices and 3 β-strands. They are associated to very high *CI* values (2.49 and 3.75 respectively). Next cutting events cut so the central PUs into 3 PUs and at the end finally into 8 PUs.

Depending on the purpose of the research, the final number PUs and / or their lengths and contents can be different. Here some PUs are only 20 residues long and associated to low CI, *e.g.*, PU 60-80 has null *CI*. It is so interesting to come back at previous cutting events. Our web server allows coming back to each cutting events. It is also always possible to change the options concerning length, R^2 values, etc.

Conclusion

The three-dimensional structure is the core of protein functions and is mainly determined by its amino acid sequence. Nonetheless, the protein folding is not completely understood (Clark 2008). Several models have been proposed for protein folding, *e.g.*, the

framework model (Ptitsyn and Rashin 1975; Udgaonkar and Baldwin 1988), the diffusion-collision model (Karplus and Weaver 1994), the hydrophobic collapse model (Rackovsky and Scheraga 1977) or the nucleation and growth mechanism (Fersht 1997). George Rose proposed a simple hierarchical model (Rose 1979), which assemblies small units in a hierarchical manner (Lesk and Rose 1981; Baldwin and Rose 1999a; b; Haspel et al. 2003a; b) coupled with the hydrophobic effect as the driving force (Dill 1985; Dill and Chan 1997). It leads to the construction of protein domains and complete folds.

Analyzing protein structures in terms of protein domains has been a long and fruitful research area for a long time. Many different approaches have been proposed (Holm and Sander 1994; Siddiqui and Barton 1995; Swindells 1995; Holm and Sander 1998; Wernisch et al. 1999; Xu et al. 2000; Anselmi et al. 2001; Dengler et al. 2001; Siddiqui et al. 2001; Alexandrov and Shindyalov 2003; Guo et al. 2003; Emmert-Streib and Mushegian 2007; Joshi 2007; Zhou et al. 2007). They are based on numerous processes and algorithms. Most of them had initially an available website, but surprisingly at the time of this review none is functional.

Protein domains are also evolutionary units of proteins. The prediction of protein domains from sequence information can improve tertiary structure prediction (Chivian et al. 2003) and enhance protein function annotation (Holland et al. 2006), but domains also been used to help structure determination (Campbell and Downing 1994), guide protein engineering (Guerois and Serrano 2001) and mutagenesis (Nielsen and Yamada 2001). Hence, some approaches have been proposed to predict protein structural domains from the sole knowledge of the sequence. For instance, DOMAC (http://www.bioinfotool.org/domac.html) is a hybrid domain prediction web service integrating template-based and *ab initio* methods (Cheng 2007). Its template-based method is accurate enough for guiding protein structure prediction, structure determination, function

annotation, mutagenesis analysis and protein engineering. Other are more specialized as OPUS-Dom (Wu et al. 2009), a *de novo* method for predicting protein domain boundaries. Its methodology is based on a coarse-grained folding method, which constructs low-resolution structural models from a target sequence by folding a chain of vectors representing the predicted secondary-structure elements.

Analyzing protein structures in terms of small compact protein units is a less common research. As we shown only two methods are available at this time, DIAL and Protein Peeling. Interest of DIAL is the proposition of potential alternative splitting events. Interest of Protein Peeling is the availability of numerous options allowing a real expertise of the protein structure. Moreover, visualization tools allow a direct analysis of the cutting through JMol applet while PyMol script and accompanying Figures are of great quality. In the same way, the different Figures generated through R software, (i) dendogram showing the entire process of splitting, (ii) the presentation of PUs with secondary structures and (iii) contact map with delineation of PUs, are an efficient representation. All these points make the Protein Peeling web server a unique tool to analyze protein structures.

In the same way, our database of pre-cutting proteins provides useful materials for further analysis on structure, size, composition in amino acid and secondary structures of protein units. Such experiments open the way to other ambitious development like construction of three dimensional structures of proteins with protein units as it has been shown with similar approaches (Haspel et al. 2003a; Inbar et al. 2003). As shown, Protein Units a valuable tools to understand protein folding, predict protein structure, identify structural domains. Futures developments will concern mainly the use of PUs for classification and for prediction purposes.

Acknowledgments

This work was supported by grants from the Ministère de la Recherche, Université Paris Diderot – Paris 7, National Institute for Blood Transfusion (INTS) and the National Institute for Health and Medical Research (INSERM).

Figures

Figure 1. *DIAL web server*. (a) the entrance page of DIAL where the use upload the PDB file, (b) the result of DIAL cutting process of dialkylglycine decarboxylase (PDB code 1ZOB), (c) visualization of the cutting with Molscript software (Kraulis 1991).

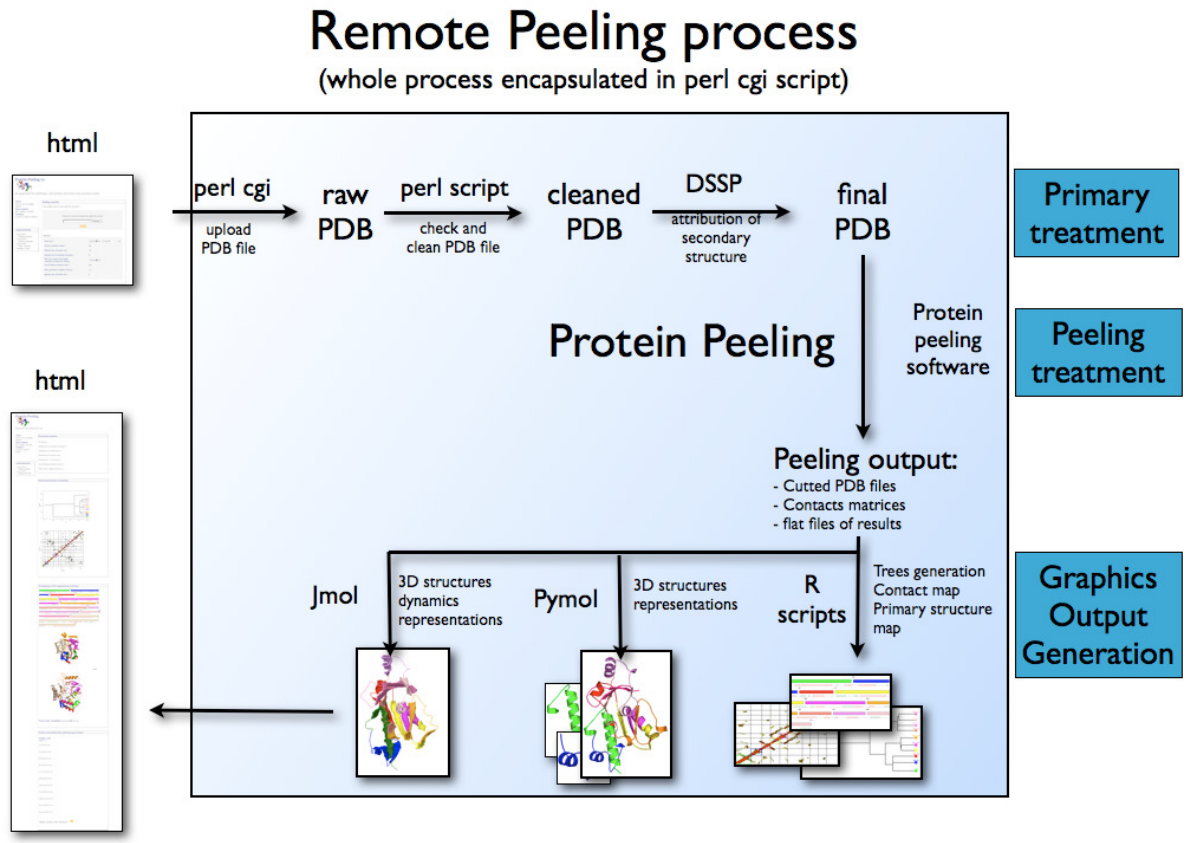

Figure 2. *Principle of PP2 web server*. All the different steps of the process are presented with the actions done and the language / software.

Figure 3. *Web page of Protein Peeling 2*. On the left part is given the different information needed (methods and contacts). On the right part, the pdb file is the only obligation; all the options are given by default and could be changed.

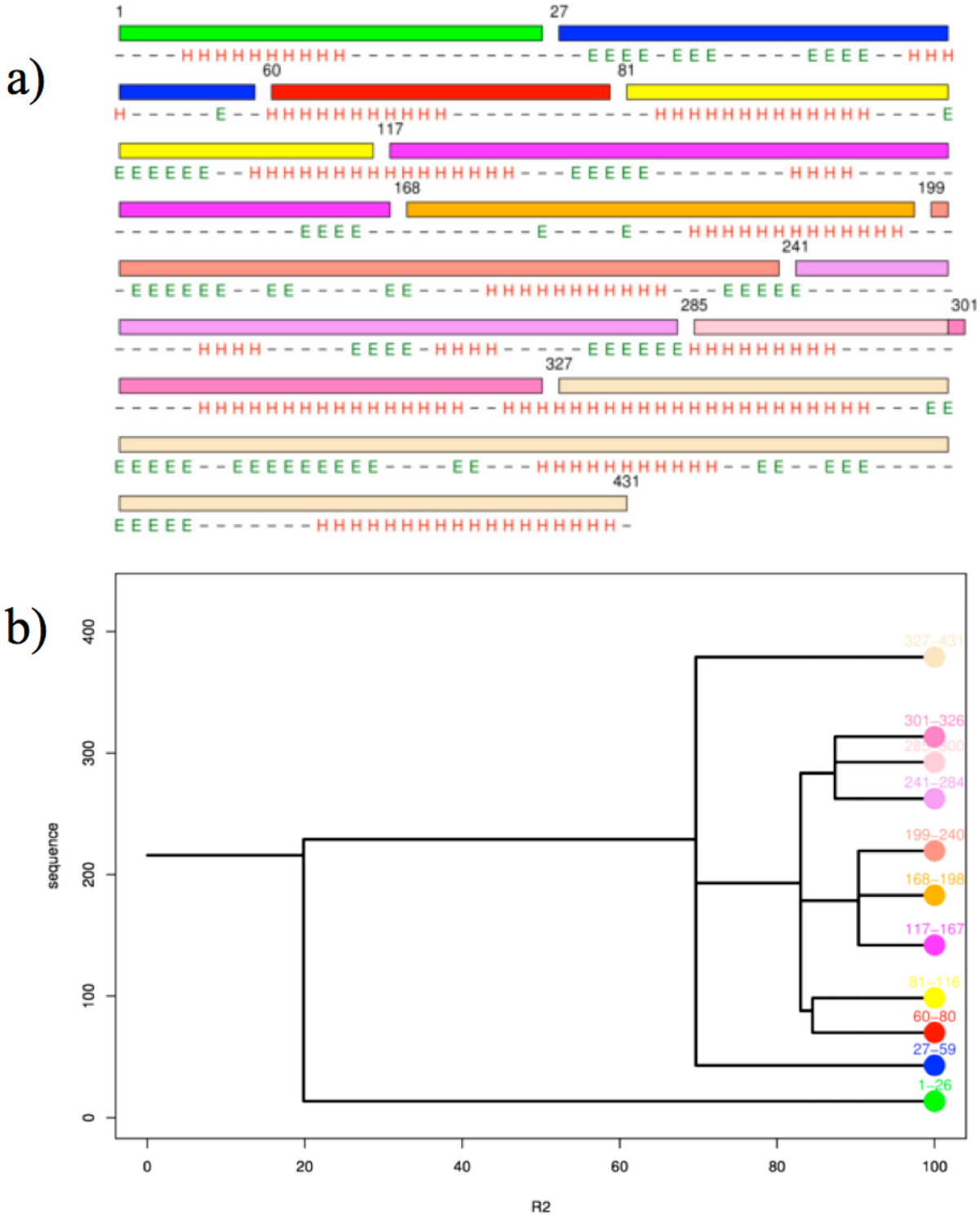

Figure 4. *PP2 cutting process of dialkylglycine decarboxylase*. (a) Representation of the protein sequence with delineation of the different PUs in different colors. Are also given the secondary structure assignment done by DSSP (Kabsch and Sander 1983). (b) Dendograms of the PP2 cutting. Is shown for each R^2 value the number of generated PUs (Ihaka and Gentleman 1996).

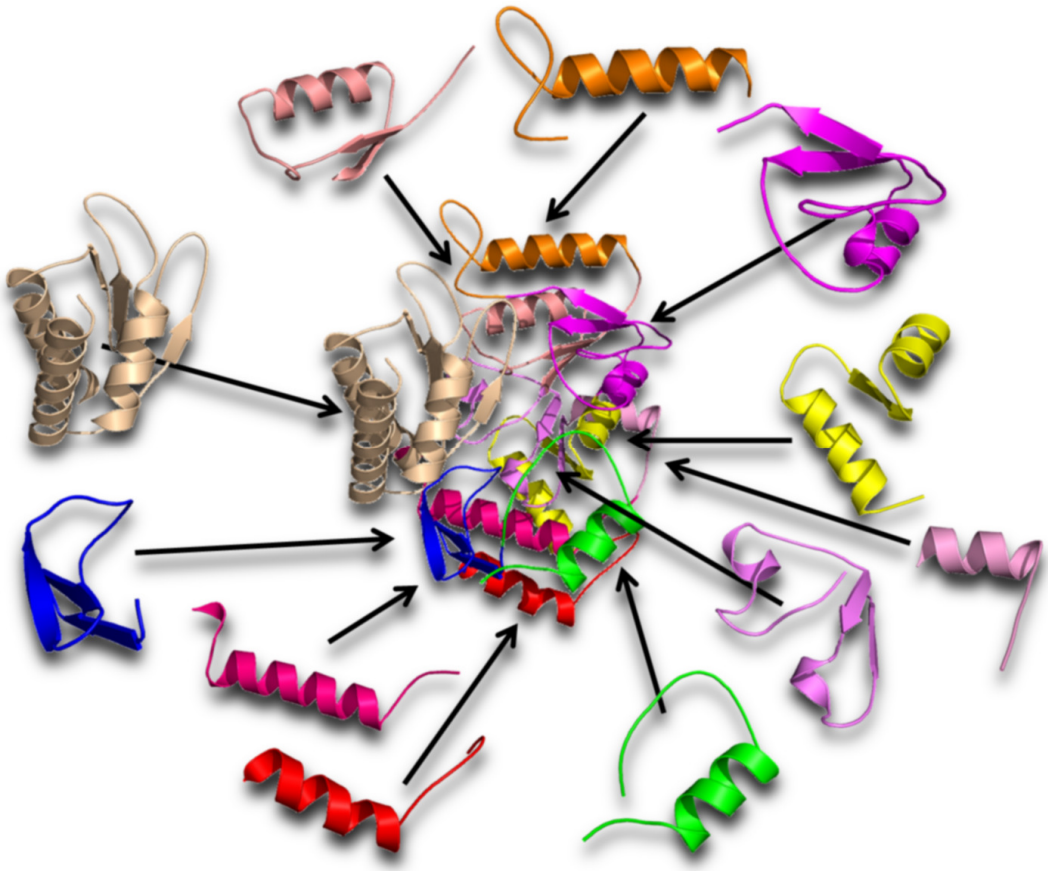

Figure 5. *PP2 cutting process of dialkylglycine decarboxylase*. Are shown all the different PUs generated for a very high R^2 value.

References

Alexandrov, N., and Shindyalov, I. 2003. PDP: protein domain parser. *Bioinformatics* **19:** 429-430.

Anfinsen, C.B., Haber, E., Sela, M., and White, F.H., Jr. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A* **47:** 1309-1314.

Anselmi, C., Bocchinfuso, G., Scipioni, A., and De Santis, P. 2001. Identification of protein domains on topological basis. *Biopolymers* **58:** 218-229.

Baldwin, R.L., and Rose, G.D. 1999a. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem Sci* **24:** 26-33.

Baldwin, R.L., and Rose, G.D. 1999b. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem Sci* **24:** 77-83.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Campbell, I.D., and Downing, A.K. 1994. Building protein structure and function from modular units. *Trends Biotechnol* **12:** 168-172.

Cheng, J. 2007. DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Res* **35:** W354-356.

Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E., Bonneau, R., Rohl, C.A., and Baker, D. 2003. Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53 Suppl 6:** 524-533.

Clark, A.C. 2008. Protein folding: Are we there yet? *Archives of Biochemistry and Biophysics* **469:** 1-3.

DeLano, W.L.T. 2002. The PyMOL Molecular Graphics System *DeLano Scientific, San Carlos, CA, USA*. http://www.pymol.org.

Dengler, U., Siddiqui, A.S., and Barton, G.J. 2001. Protein structural domains: analysis of the 3Dee domains database. *Proteins* **42:** 332-344.

Dill, K.A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* **24:** 1501-1509.

Dill, K.A., and Chan, H.S. 1997. From Levinthal to pathways to funnels. *Nat Struct Biol* **4:** 10-19.

Emmert-Streib, F., and Mushegian, A. 2007. A topological algorithm for identification of structural domains of proteins. *BMC Bioinformatics* **8:** 237.

Etchebest, C., Benros, C., Hazout, S., and de Brevern, A.G. 2005. A structural alphabet for local protein structures: improved prediction methods. *Proteins* **59:** 810-827.

Faure, G., Bornot, A., and de Brevern, A.G. 2009. Analysis of protein contacts into Protein Units. *Biochimie* **91:** 876-887.

Fersht, A. 1997. Nucleation mechanism in protein folding. *Curr. Opin. Struct. Biol* **7:** 3-9.

Gelly, J.C., de Brevern, A.G., and Hazout, S. 2006a. 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments. *Bioinformatics* **22:** 129-133.

Gelly, J.C., Etchebest, C., Hazout, S., and de Brevern, A.G. 2006b. Protein Peeling 2: a web server to convert protein structures into series of protein units. *Nucleic Acids Res* **34:** W75-78.

Go, M. 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291:** 90-92.

Guerois, R., and Serrano, L. 2001. Protein design based on folding models. *Curr Opin Struct Biol* **11:** 101-106.

Guo, J.T., Xu, D., Kim, D., and Xu, Y. 2003. Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res* **31:** 944-952.

Haspel, N., Tsai, C.J., Wolfson, H., and Nussinov, R. 2003a. Hierarchical protein folding pathways: a computational study of protein fragments. *Proteins* **51:** 203-215.

Haspel, N., Tsai, C.J., Wolfson, H., and Nussinov, R. 2003b. Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci* **12:** 1177-1187.

Hazout, S. 2007. Entropy-derived measures for assessing the accuracy of N-state prediction algorithms. In *Recent Advances in Structural Bioinformatics*. (ed. A.G. de Brevern), pp. 395-417. Research signpost, Trivandrum, India.

Heger, A., Wilton, C.A., Sivakumar, A., and Holm, L. 2005. ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res* **33:** D188-191.

Holland, T.A., Veretnik, S., Shindyalov, I.N., and Bourne, P.E. 2006. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol* **361:** 562-590.

Holm, L., and Sander, C. 1994. Parser for protein folding units. *Proteins* **19:** 256-268.

Holm, L., and Sander, C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* **25:** 231-234.

Holm, L., and Sander, C. 1998. Dictionary of recurrent domains in protein structures. *Proteins* **33:** 88-96.

Ihaka, R., and Gentleman, R. 1996. R: a language for data analysis and graphics. *J Comput Graph Stat* **5:** 299-314.

ImageMagick. http://www.imagemagick.org.

Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H.J. 2003. Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics* **19 Suppl 1:** i158-168.

Jacob, E., and Unger, R. 2007. A tale of two tails: why are terminal residues of proteins exposed? *Bioinformatics* **23:** e225-230.

Janin, J., and Wodak, S.J. 1983. Structural domains in proteins and their role in the dynamics of protein function. *Prog Biophys Mol Biol* **42:** 21-78.

JMol. http://jmol.sourceforge.net/.

Joshi, R.R. 2007. A Decade of Computing to Traverse the Labyrinth of Protein Domains. *Current Bioinformatics* **2:** 113-131.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Karplus, M., and Weaver, D.L. 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci* **3:** 650-668.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181:** 662-666.

Kraulis, P. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24:** 946-950

Kundu, S., Sorensen, D.C., and Phillips, G.N., Jr. 2004. Automatic domain decomposition of proteins by a Gaussian Network Model. *Proteins* **57:** 725-733.

Lesk, A.M., and Rose, G.D. 1981. Folding units in globular proteins. *Proc Natl Acad Sci U S A* **78:** 4304-4308.

Matthews, B. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405:** 442-451.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247:** 536-540.

Nielsen, P.K., and Yamada, Y. 2001. Identification of cell-binding sites on the Laminin alpha 5 N-terminal domain by site-directed mutagenesis. *J Biol Chem* **276:** 10906-10912.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH--a hierarchic classification of protein domain structures. *Structure* **5:** 1093-1108.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., and North, A.C. 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* **185:** 416-422.

Ponting, C.P., and Russell, R.R. 2002. The natural history of protein domains. *Annu Rev Biophys Biomol Struct* **31:** 45-71.

Ptitsyn, O.B., and Rashin, A.A. 1975. A model of myoglobin self-organization. *Biophys Chem* **3:** 1-20.

Pugalenthi, G., Archunan, G., and Sowdhamini, R. 2005. DIAL: a web-based server for the automatic identification of structural domains in proteins. *Nucleic Acids Res* **33:** W130-132.

Rackovsky, S., and Scheraga, H.A. 1977. Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. *Proc Natl Acad Sci U S A* **74:** 5248-5251.

Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34:** 167-339.

Rose, G.D. 1979. Hierarchic organization of domains in globular proteins. *J Mol Biol* **134:** 447-470.

Rossman, M.G., and Liljas, A. 1974. Letter: Recognition of structural domains in globular proteins. *J Mol Biol* **85:** 177-181.

Siddiqui, A.S., and Barton, G.J. 1995. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci* **4:** 872-884.

Siddiqui, A.S., Dengler, U., and Barton, G.J. 2001. 3Dee: a database of protein structural domains. *Bioinformatics* **17:** 200-201.

Sowdhamini, R., and Blundell, T.L. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci* **4:** 506-520.

Sowdhamini, R., Rufino, S.D., and Blundell, T.L. 1996. A database of globular protein structural domains: clustering of representative family members into similar folds. *Fold Des* **1:** 209-220.

Swindells, M.B. 1995. A procedure for detecting structural domains in proteins. *Protein Sci* **4:** 103-112.

Taylor, W.R. 2007. Evolutionary transitions in protein fold space. *Curr Opin Struct Biol* **17:** 354-361.

Tsai, C.J., and Nussinov, R. 1997. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci* **6:** 24-42.

Udgaonkar, J.B., and Baldwin, R.L. 1988. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. *Nature* **335:** 694-699.

Wernisch, L., Hunting, M., and Wodak, S.J. 1999. Identification of structural domains in proteins by a graph heuristic. *Proteins* **35:** 338-352.

Wetlaufer, D.B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A* **70:** 697-701.

Wetlaufer, D.B. 1981. Folding of protein fragments. *Adv Protein Chem* **34:** 61-92.

Wodak, S.J., and Janin, J. 1981. Location of structural domains in protein. *Biochemistry* **20:** 6544-6552.

Wu, Y., Dousis, A.D., Chen, M., Li, J., and Ma, J. 2009. OPUS-Dom: applying the folding-based method VECFOLD to determine protein domain boundaries. *J Mol Biol* **385:** 1314-1329.

Xu, Y., Xu, D., and Gabow, H.N. 2000. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* **16:** 1091-1104.

Zhou, H., Xue, B., and Zhou, Y. 2007. DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. *Protein Sci* **16:** 947-955.